YOSHUA BENGIO, Université de Montréal, and the Founder and Scientific Director of Mila – Quebec AI Institute

Mathematical challenges towards safe AI

Advances in algorithms and computational capabilities of AI systems based on deep learning have been impressive and herald possibly disruptive transformations in coming years and decades, with great potential for both benefits and risks for humanity. The three winners of the Turing award for deep learning (2018) expect that broad human-level capabilities are likely to be achieved within just a few years or decades and industry is investing billions of dollars per month which are likely to accelerate this process. However, we do not yet know how to design provably safe and controllable AI systems, i.e., systems that behave as we intend. This misalignment could threaten democracy, national security and possibly our collective future either due to malicious actors or a loss of control to runaway AIs. Worse, arguments have been made suggesting that the state-of-the-art AI methodology, based on reinforcement learning, would yield less and less safety as computational power increases. This presentation will argue that there may be a way to design AI systems with probabilistic safety guarantees that improve as we increase the computational capabilities of the underlying neural networks. This would rely on efficient and amortized Bayesian inference in learned causal models, designing AI systems inspired by how scientists and mathematicians come up with theories that are compatible with reason and observed evidence.